# On The Mel-scaled Cepstrum

H.P. Combrinck and E.C. Botha
Department of Electrical and Electronic Engineering
University of Pretoria
Pretoria
South Africa
E-mail: rikus@suntiger.ee.up.ac.za

*Abstract*— **The mel-scaled cepstrum is a signal representation scheme used in the analysis of speech signals. Due to its reported superior performance, especially under adverse conditions, it is becoming an increasingly popular choice as feature extraction front end to spoken language systems. Having evolved over a period of more than fifty years, the mel-scaled cepstrum owes part of its heritage to the pattern recognition community and part to perceptual and acoustical research. It represents a good trade-off between computational efficiency and perceptual considerations. Unfortunately, maybe because of its hybrid nature, the literature tends to be vague on the implementation details of mel-scaled cepstrum algorithms. In this paper we clarify some of the issues regarding the algorithm and its implementation. Our investigation also serves to expose some fundamental flaws remaining in the established approach to speech signal feature extraction.**

## I. Introduction

THE pre-processing and feature extraction stages of a pattern recognition system serves as an interface between the real world and a classifier operating on an idealised model of reality. Information that is discarded in this stage is forever lost; conversely, noise that is accepted will degrade the performance of the classifier stage that is typically sensitive to complexity in the data. The signals that spoken language systems have to deal with is unique in the sense that it is generated by a biological system, for a biological system. Human speech is the evolutionary product of the vocal and auditory systems and not the other way around. The result shows a distinct lack of engineering common sense. As a matter of fact, psychophysical studies over the last number of decades tend to leave us with the uncomfortable feeling that the world perceived through our senses is rather different from the one that we measure with our instruments. We will now consider some revealing aspects of human auditory preception and then examine the mel-scaled cepstrum algorithm in order to draw some conclusions.

## II. Peculiarities of the Human Auditory System

A pure tone is uniquely defined by its intensity and frequency. The perceptual counterparts of these quantities are termed loudness and pitch respectively. Pitch is difficult to define. Mostly we agree that pure tones can be ordered in such a way that one tone is 'higher' or 'lower' than another. Pitch is the criterion that we use to make such decisions. Like loudness, it is a complex, non-linear function of both frequency and intensity. Stevens, Volkmann and Newman defined the mel scale, which relates pitch to frequency as depicted in Fig. 1 [1]. It was later refined by Stevens and Volkmann in their classical paper [2]. The form of the curve was determined by perceptual experiments designed to find a linear relation among perceived pitches. A pitch of 2000 mels is therefore subjectively 'twice as high' as a pitch of 1000 mels. The numeric range of the mel scale and its relation to sound intensity was fixed by defining a 40dB tone with a frequency of 1000Hz as having a pitch of 1000 mels. We fitted a curve on Stevens' and Volkmann's original data to obtain (1) where $f$ denotes frequency in Hertz and $\nu$, pitch in mels.

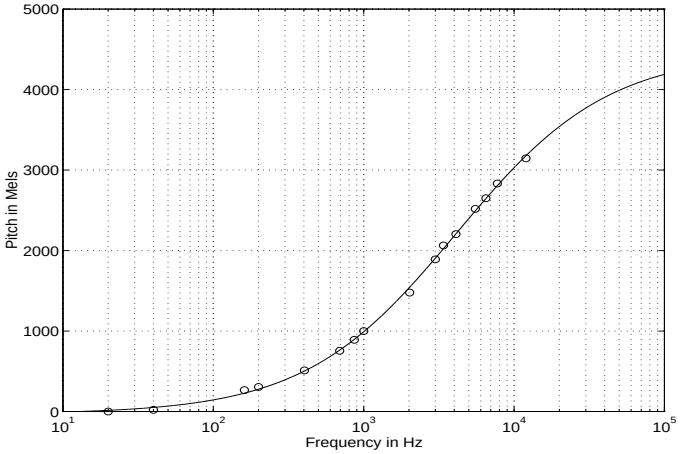


Fig. 1. The mel scale

$$\nu(f) = \frac{4491.7}{(1 + \exp(7.1702 - 1.9824 \log_{10}(f))} - 30.360 \tag{1}$$

Loudness is a psychological term used to describe the magnitude of an auditory sensation and is a function of intensity and frequency, as well as a number of psychological factors like fatigue, attention and alertness. Fletcher and Munson investigated and defined loudness [3]. They extended their work in [4], where they also addressed masking, an interesting auditory phenomenon: the threshold at which a tone can be perceived is raised when heard in the presence of another tone (or band of noise). The effect of a sound on the auditory system persists for milliseconds and therefore the perception of a sound depends on its context. This is called forward masking. Fletcher found that a pure tone is masked essentially only by noise components within a

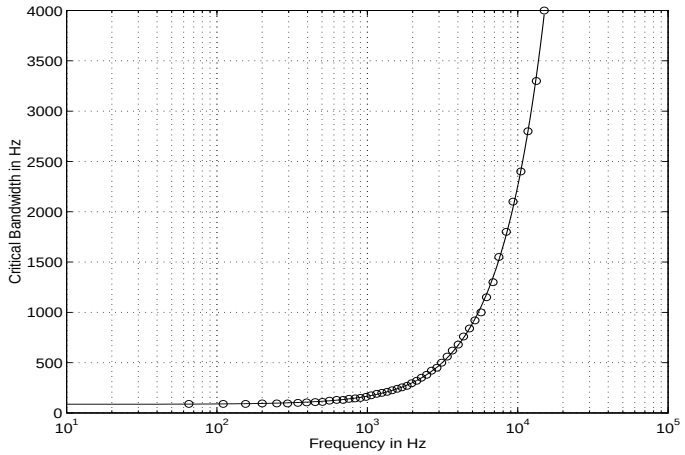


Fig. 2. Critical bandwidth as a function of frequency

certain narrow band centred at the frequency of the tone [5]. Differential pitch sensitivity, the smallest detectable change in frequency (also a strong function of sound intensity and frequency), is closely linked to these bands [6]. As a matter of fact, a number of perceptual phenomena seem to indicate that there exist what came to be called critical bands [7], [8], [9], [10]. The bandwidth of critical bands increase with frequency as shown in Fig. 2. Equations (2) and (3) describe the the critical band cutoff frequencies.

$$\text{freq}_{\text{crit}_{\text{low}}}(f) = 1.3056 f^{0.95987} - 64.193 \tag{2}$$

$$\text{freq}_{\text{crit}_{\text{high}}}(f) = 0.70616 f^{1.0497} + 81.288 \tag{3}$$

The bark scale, another pitch scale that corresponds closely in form to the mel scale, is defined in terms of critical bands [11], [12].

Spoken language processing research is conducted by engineers that are slowly drifting towards auditory modelling and psychologists, physiologists and linguists that are moving towards digital signal processing and pattern recognition. Steadily increasing processing power gradually moved the point of optimal performance towards the more complex and computationally expensive modelling approach.

Engineers were lured into the world of auditory perception mainly through their attempts to optimise telephone systems. A classic study on the intelligibility of speech can be found in [13]. As the performance of digital computers exploded, it opened the world to speech recognition experiments. Following Bridle and Brown [14], Mermelstein [15] investigated the ability of the mel-scaled cepstrum to distinguish between similar sounding consonants. In a later publication Davies and Mermelstein [16] found the mel-scaled cepstrum to be significantly superior to four other feature extraction front ends in a syllable-oriented speaker dependent, continuous speech recognition task. A recent study compared the mel-scaled cepstrum to two feature extraction front ends based on auditory models on a speaker dependent word recognition task [17]. It was shown that the more complex front ends provided little improvement in performance (a difference of 0.6 to 4 percentage points in error rate) to compensate for increased complexity and processing time ($\frac{1}{3}$ real time as opposed to respectively, 40 and 120 times real time). In addition it was shown that the mel-scaled cepstrum approach significantly outperforms a traditional LPC-based front end. These results were extended to a (male) speaker independent, continuous task [18]. We now present an algorithm to calculate the mel-scaled cepstrum. This algorithm has been reconstructed from [18], [19], [15], [17] and our own experience. We believe it to be in a format that will easily translate into a high-level programming language implementation.

## III. An Algorithm

Let the $N$-sample speech signal be

$$\mathbf{x} = x_0, \ldots, x_{N-1} \tag{4}$$

### A. Preemphasis

The speech signal is preemphasized to compensate for spectral tilt (i.e. $S'(w) = S(w).w^a$). This is a high-pass filtering operation and can be executed in either the time or frequency domain. The filter in the time-domain is of the form

$$x_i = x_i - a x_{i-1}, \quad 0.9 \le a \le 1.0 \tag{5}$$

where the parameter $a$ is not critical and is usually taken to be 0.95.

### B. Normalisation

The maximum signal amplitude is normalised to one.

$$x_i = \frac{x_i}{\max_{j=0,\ldots,N-1} \left| x_j \right|}, \quad i = 0, \ldots, N-1 \tag{6}$$

### C. Blocking

The filtered, normalised signal is broken into $M$ overlapping frames and stored in an $M \times W$ matrix $\mathbf{Y}$ with its rows $\mathbf{y}_i$ representing the frames. $V$ is the step size and $W$ the frame size. Frame size usually range from 10ms to 20ms and step size between 20 and 50 percent of frame size.

$$y_{ij} = x_{Vi+j}, \quad j = 0, \ldots, W-1, \quad i = 0, \ldots, M-1 \tag{7}$$

### D. Windowing

Each frame is multiplied with a window function to minimise signal discontinuities in the time domain and the resulting spectral artifacts.

$$y_{ij} = y_{ij} w_j, \quad j = 0, \ldots, W-1, \quad i = 0, \ldots, M-1 \tag{8}$$

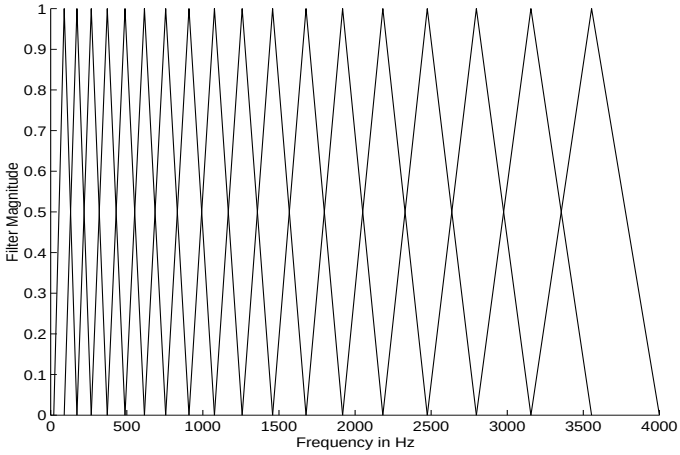


Fig. 3. A mel scale filter bank

The Hamming window, described by (9), is a popular choice.

$$w_j = 0.54 - 0.46\cos\left(\frac{2\pi j}{W-1}\right), \quad j = 0,\ldots,W-1 \quad (9)$$

### *E. Power Spectrum*

The power spectrum of each window is calculated and represented by the the $M \times U$ matrix $\mathbf{S}$. $W$ (and $U$) will be constrained by the FFT algorithm in a practical implementation. We used a prime factor FFT which gives more freedom in the choice of $W$ than the standard radix-2 algorithms. Still, $W$ needs to be one of a limited set of integers that will in general not be the same as the number determined by the choice of frame size. To work around this, $\mathbf{y}_i$ can be zero-padded or the frame size can be adjusted to coincide with a valid number.

$$\mathbf{s}_i = |\mathrm{fft}(\mathbf{y}_i)|^2, \quad i = 0,\ldots,M-1 \quad (10)$$

### *F. Mel Filter Bank*

The mel filter bank consists of overlapping triangular filters with the cutoff frequencies determined by the centre frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale. This arrangement (depicted in Fig. 3) results in a logarithmic spacing on the frequency scale with bandwidths roughly corresponding to the critical bandwidth curve. This models the differential pitch sensitivity of the ear. If one takes $f_{\min}$ as 20Hz (0 mels) and $f_{\max}$ as half the sampling rate, then the mel filter bank is defined by equations (11) to (13) and represented by the $K \times U$ matrix $\mathbf{F}$. $f_c$ is the centre frequency of a filter. The low and high cutoff frequencies, $f_l$ and $f_h$, are the centre frequencies of the two adjacent filters. The number of filters, $K$, is usually between 13 and 24. Care must be taken that it is not too large, since crowding the filters will result in poor frequency resolution at low frequencies. Let

$$\gamma = \frac{W-1}{f_{\max} - f_{\min}} \quad (11)$$

$$\begin{aligned} I_l &= \gamma(f_l - f_{\min}) \\ I_c &= \gamma(f_c - f_{\min}) \\ I_r &= \gamma(f_h - f_{\min}) \end{aligned} \quad (12)$$

then

$$f_{ij} = \begin{cases} \frac{1}{f_c - f_l}\left(\frac{j}{\gamma} + f_{\min} - f_l\right) & \text{if } \lceil I_l \rceil \le j \le \lfloor I_c \rfloor \\ 1 + \frac{1}{f_c - f_h}\left(\frac{j}{\gamma} + f_{\min} - f_c\right) & \text{if } \lceil I_c \rceil \le j \le \lfloor I_r \rfloor \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

An approximation to the mel scale that is frequently used for the filter bank, is to have a number of linearly spaced filters with equal bandwidth under 1000Hz and then logarithmically spaced filters above 1000Hz where the centre frequency of each filter is 1.1 times the preceding centre frequency [17], [18].

### *G. Log Energy Filter Coefficients*

Now, one of the confusing aspects of the mel-scaled cepstrum is that the mel filter bank is not really a filter. In stead of just weighting each point of the spectrum with the filter weights, one calculates a type of inner product and find what is probably best classified as an energy coefficient for each filter. To compensate for the increasing bandwidths of the filters, the energies are normalised by (15). This part of the processing is completed by taking the logarithm of each energy coefficient in a crude attempt to model the non-linear intensity-loudness relationship which is logarithmic in nature. These operations result in the $M \times K$ matrix $\mathbf{P}$.

$$p_{ij} = \log_{10}\left(\frac{1}{A_j}\sum_{k=0}^{U-1} s_{ik} f_{jk}\right), \quad \begin{matrix} j = 0,\ldots,K-1 \\ i = 0,\ldots,M-1 \end{matrix} \quad (14)$$

where

$$A_j = \sum_{k=0}^{U-1} f_{jk} \quad (15)$$

### *H. Inverse Discrete Cosine Transform*

The inverse cosine transform is used to orthogonalise the filter energy vectors. It has been suggested to be an efficient approximation to the optimal Karhunen-Loeve transform [20], [21], [22]. Because of this orthogonalisation step, the information of the filter energy vector is compacted into the first number of components and we can shorten the vector to $L$ components, resulting in the $M \times L$ matrix $\mathbf{Q}$.

$$q_{ij} = \frac{1}{K}\sum_{k=0}^{K-1} p_{ik}\cos\left((k-0.5)\frac{\pi j}{L}\right), \quad \begin{matrix} j = 0,\ldots,L-1 \\ i = 0,\ldots,M-1 \end{matrix} \quad (16)$$

$L$ is chosen to be less than $K$, usually somewhere between 9 and 15.

### I. Incorporation of Dynamic Features

It has been found that including the first and second derivatives of the log energy vector significantly improves the performance of mel-scaled cepstrum-based systems [18], [17]. These are referred to as the delta and delta-delta cepstra. Since we are dealing with discrete data, it is advisable to calculate the derivatives on smoothed data. This issue is considered in detail in [23]. The first mel-scaled cepstral coefficient represent the mean energy in each frame and is usually dropped. The delta-cepstrum and occasionally the delta-delta cepstrum is concatenated to the mel-scaled cepstrum to form one long vector. This then constitutes a mel-scaled cepstrum feature vector.

## IV. Discussion

From a engineering point of view, the mel-scaled cepstrum is an efficient algorithm because it is performed mainly in the frequency domain and we can use the FFT. The inverse cosine transform is an efficient dimension reduction technique as well. From a perceptual point of view, the mel-scaled cepstrum takes into account the non-linear nature of pitch perception (the mel scale) as well as loudness perception (the log operation). It also models critical bandwidth as far as differential pitch sensitivity is concerned (the mel scale). The delta-cepstrum incorporates dynamic information. The mel-scaled cepstrum does not model (static) masking. It does not model forward (dynamic) masking and there is no feedback between higher level processing (the classifier stage) and feature extraction.

## V. Conclusion

We have discussed the mel-scaled cepstrum feature extraction algorithm in the context of significant human auditory phenomena and found it to be a good engineering solution compared to other standard approaches. Still, despite the fact that spoken language systems function reasonably well and are continually improving, they still lag far behind human performance, especially under adverse conditions. We feel that this is due to the fact that current systems largely ignore the dynamic nature of the auditory system. Specifically, masking and forward masking are very important concepts that need to be integrated in an efficient way into spoken language feature extraction systems [24]. As processing power increase we may find it rewarding to move to more advanced auditory models.

## References


[1] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, Jan. 1937.

[2] S. S. Stevens and J. Volkmann, "The relation of pitch to frequency: A revised scale," *American Journal of Psychology*, vol. 53, pp. 329–353, July 1940.

[3] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *Journal of the Acoustical Society of America*, vol. 5, pp. 82–108, Oct. 1933.

[4] H. Fletcher and W. A. Munson, "Relation between loudness and masking," *Journal of the Acoustical Society of America*, vol. 9, pp. 1–10, July 1937.

[5] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, pp. 47–65, Jan. 1940.

[6] E. G. Shower and R. Biddulph, "Differential pitch sensitivty of the ear," *Journal of the Acoustical Society*, vol. 3, pp. 275–287, Oct. 1931.

[7] S. S. Stevens, "Calculation of the loudness of complex noise," *Journal of the Acoustical Society of America*, vol. 28, pp. 807–832, Sept. 1956.

[8] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical bandwidth in loudness summation," *Journal of the Acoustical Society of America*, vol. 29, pp. 548–557, May 1957.

[9] T. H. Schafer, R. S. Gales, C. A. Shewmaker, and P. O. Thompson, "The frequency selectivity of the ear as determined by masking experiments," *Journal of the Acoustical Society of America*, vol. 22, pp. 490–496, July 1950.

[10] J. E. Hawkins, Jr. and S. S. Stevens, "The masking of pure tones and of speech by white noise," *Journal of the Acoustical Society of America*, vol. 22, pp. 6–13, Jan. 1950.

[11] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *Journal of the Acoustical Society of America*, vol. 33, p. 248, Feb. 1961.

[12] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *Journal of the Acoustical Society of America*, vol. 68, pp. 1523–1525, Nov. 1980.

[13] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *Journal of the Acoustical Society of America*, vol. 19, pp. 90–119, Jan. 1947.

[14] J. S. Bridle and M. D. Brown, "An experimental automatic word recognition system," Tech. Rep. JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England, 1974. Referenced in [15].

[15] P. Mermelstein, "Distance measures for speech recognition - psychological and instrumental," in *Pattern Recognition and Artificial Intelligence* (C. H. Chen, ed.), pp. 374–388, Academic Press, Inc., 1976.

[16] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

[17] C. R. Jankowski Jr., H.-D. H. Vo, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 286–293, July 1995.

[18] S. Sandhu, "A comparative study of mel cepstra and EIH for phone classification under adverse conditions," Master's thesis, Massachusetts Institute of Technology, Feb. 1995.

[19] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, section 4.5.6, pp. 183–190. In *Prentice Hall Signal Processing Series* [25], 1993.

[20] N. Merhav and C.-H. Lee, "On the asymptotic statistical behaviour of empirical cepstral coefficients," *IEEE Transactions on Signal Processing*, vol. 41, pp. 1990–1993, May 1993. Referenced in [24].

[21] L. C. W. Pols, *Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words*. PhD thesis, Free University, Amsterdam, 1966. Referenced in [15], [17].

[22] J. F. Blinn, "What's the deal with the DCT?," *IEEE Computer Graphics and Applications*, vol. 13, pp. 78–83, July 1993.

[23] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, section 4.6, pp. 194–200. In *Prentice Hall Signal Processing Series* [25], 1993.

[24] B. P. Strope, "A model of dynamic auditory perception and its application to robust speech recognition," Master's thesis, University of California, 1995.

[25] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, Prentice-Hall, Inc., 1993.